



Multiple point statistical simulation using uncertain (soft) conditional data

Hansen, Thomas Mejer; Vu, Le Thanh; Mosegaard, Klaus; Cordua, Knud Skou

Published in:
Computers & Geosciences

DOI:
[10.1016/j.cageo.2018.01.017](https://doi.org/10.1016/j.cageo.2018.01.017)

Publication date:
2018

Document version
Peer reviewed version

Document license:
[CC BY-NC-ND](#)

Citation for published version (APA):
Hansen, T. M., Vu, L. T., Mosegaard, K., & Cordua, K. S. (2018). Multiple point statistical simulation using uncertain (soft) conditional data. *Computers & Geosciences*, 114, 1-10.
<https://doi.org/10.1016/j.cageo.2018.01.017>

Multiple Point Statistical simulation using uncertain (soft) conditional data

Thomas Mejer Hansen^a, Le Thanh Vu^b, Klaus Mosegaard^a, Knud Skou
Cordua^a

^a*Niels Bohr Institute, University of Copenhagen, Denmark*

^b*I-GIS, Risskov, Denmark*

^c*This is the final manuscript accepted for publication in Computers and Geosciences,
<https://doi.org/10.1016/j.cageo.2018.01.017>*

Abstract

Geostatistical simulation methods have been used to quantify spatial variability of reservoir models since the 80s. In the last two decades, state of the art simulation methods have changed from being based on covariance-based 2-point statistics to multiple-point statistics (MPS), that allow simulation of more realistic Earth-structures. In addition, increasing amounts of geo-information (geophysical, geological, etc.) from multiple sources are being collected. This pose the problem of integration of these different sources of information, such that decisions related to reservoir models can be taken on an as informed base as possible. In principle, though difficult in practice, this can be achieved using computationally expensive Monte Carlo methods. Here we investigate the use of sequential simulation based MPS simulation methods conditional to uncertain (soft) data, as a computational efficient alternative. First, it is demonstrated that current implementations of sequential simulation based on MPS (e.g. SNESIM, ENESIM and Direct Sampling) do not account properly for uncertain conditional information, due to a combination of using only co-located information, and a random simulation path. Then, we suggest two approaches that better account for the available uncertain information. The first make use of a preferential simulation path, where more informed model parameters are visited preferentially to less informed ones. The second approach involves using non co-located uncertain information. For different types of available data, these approaches are demonstrated to produce simulation results similar to those obtained by the general Monte Carlo based approach. These methods allow MPS simu-

lation to condition properly to uncertain (soft) data, and hence provides a computationally attractive approach for integration of information about a reservoir model.

10 *Keywords:* Multiple point statistics, Uncertain data, Data integration

11 **1. Introduction**

12 During the last 30 years a number of probabilistic based methods and
13 algorithms have been developed in the geostatistical community, that allow
14 quantification and simulation of increasingly geologically complex structural
15 variability, see e.g. Deutsch and Journel (1992); Guardiano and Srivastava
16 (1993); Strebel (2000); Remy et al. (2008); Mariethoz et al. (2010); Straub-
17 haar et al. (2011); Mariethoz and Kelly (2011); Toftaker and Tjelmeland
18 (2013); Tahmasebi et al. (2014); Mariethoz and Caers (2014).

19 State of the art simulation methods have changed from being based
20 on 2-point statistics (covariance-based statistics) to multiple-point statistics
21 (MPS), that allow simulation of more realistic Earth-structures. MPS is es-
22 pecially important used as a base for flow modeling, as traditional 2-point
23 statistics cannot adequately describe for example realistic connectivity of ge-
24 ological structures, that may have significant effect on flow properties and
25 transport, see e.g. Zinn and Harvey (2003); Renard et al. (2011). The infor-
26 mation about the expected spatial variability of the properties in a reservoir
27 model can be conveniently provided in form of a ‘training image’/‘sample
28 model’ when using MPS. Using such a training image, several methods exist
29 for simulation of multiple realizations of reservoir models that are consis-
30 tent with the spatial statistics of the training image, e.g. Guardiano and
31 Srivastava (1993); Strebel (2000); Mariethoz et al. (2010).

32 Additional information is often available from e.g. boreholes and geophys-
33 ical surveys (seismic, electromagnetic,...). Ideally, this information should be
34 combined with the geostatistical information in order to obtain a stochastic
35 reservoir model, or realizations of such a model that are consistent with all
36 available data/information.

37 Several methods have been proposed to deal with this problem of integra-
38 tion of information. Probabilistic inverse problem theory allow combining the
39 available information by characterizing (or sampling from) a posterior prob-
40 ability function that combines the information from the geostatistical model
41 that describes realistic earth models (in form of a prior probability density),

with information from data (in form of a likelihood function) (Tarantola, 2005). Using Monte Carlo sampling the posterior of any posterior probability can be sampled, as long as the prior model can be sampled, and the likelihood can be evaluated (Mosegaard and Tarantola, 1995; Hansen et al., 2008; Irving and Singha, 2010; Hansen et al., 2012; Cordua et al., 2012; Hansen et al., 2013). While such a Monte Carlo based approach can in principle deal with a large variety of very complex systems, its practical use is hampered by its very high computational demands.

Another approach is typically used in geostatistics, where available (geophysical) data are converted into ‘soft data’ about each individual model parameter. Soft data is a loosely defined term that typically refer to uncertainty and inequality constraints about specific model parameters (Journel, 1986). Most all geostatistical simulation algorithms can make use of such ‘soft’ data (Remy et al., 2008; Mariethoz and Caers, 2014). However, challenges related to using current state of the art MPS simulation algorithms conditional to other geo-information has been considered widely in the literature with respect to ground water models He et al. (2014); Koch et al. (2014); Jørgensen et al. (2015); Biver et al. (2014); Høyer et al. (2015, 2017).

In the following the use of sequential simulation based MPS sampling methods will be considered for probabilistic data integration with independent uncertain conditional data, that may be available from other sources.

First, using notation from probabilistic data integration, we formulate precisely what is implicitly assumed about ‘soft data’ in most any MPS algorithm.

Through analysis of 3 reference models, with varying density of conditional/soft data, we demonstrate that a conventional implementation of sequential simulation based MPS simulation leads to simulations that fail to generate realizations (reservoir models) consistent with the available uncertain information (soft data).

Then, we suggest two novel approaches that allow considering the information in a more correct way using direct sampling (DS, Mariethoz et al., 2010), ENESIM (Guardiano and Srivastava, 1993), and SNESIM (Strebelle, 2000). The first use as preferential simulation path, where more informed model parameters are visited preferentially to less informed ones. The second approach involves using more than only co-located uncertain data, which is typically not done for most implementations of MPS. All examples are compared to those obtained by a general Monte Carlo based approach.

80 2. Data integration using conditional geostatistical simulation - 81 Theory

Consider that a model of the subsurface is parameterized into M model parameters $\mathbf{m} = [m_1, m_2, m_3, \dots, m_M]$. Say information is available about the model parameters \mathbf{m} from N independent sources $\mathbf{I} = [I_1, I_2, \dots, I_N]$ through the probability densities $f_{I_1}(\mathbf{m})$, $f_{I_2}(\mathbf{m})$, ..., $f_{I_N}(\mathbf{m})$. Each probability distribution then represents a specific *state of information*. Tarantola and Valette (1982) and Tarantola (2005) demonstrate how these states of information can be combined through the *conjunction* of the states of information through

$$\begin{aligned} f_{\mathbf{I}}(\mathbf{m}) &= f_{I_1}(\mathbf{m}) \wedge f_{I_2}(\mathbf{m}) \wedge \dots \wedge f_{I_N}(\mathbf{m}) \\ &= \nu \mu(\mathbf{m})^{(1-N)} \prod_i^N f_{I_i}(\mathbf{m}), \end{aligned} \quad (1)$$

82 where ν represents a normalizing constant, $\mu(\mathbf{m})$ represents the homogeneous
83 probability distribution or the ‘state of total ignorance’ (Jaynes, 1968), and
84 \wedge is the operator for ‘conjunction’. Conjunction of information, as expressed
85 through (1), is derived from axioms similar to the axioms of formal logic on
86 conjunction of propositions, and the Radon-Nikodym theorem from measure
87 theory (Tarantola and Valette, 1982).

88 If a Cartesian coordinate system is used to parameterize \mathbf{m} , then the
89 homogeneous probability density function becomes a constant $\mu(\mathbf{m}) = k$
90 (Mosegaard and Tarantola, 2002), which is the case we will consider here.
91 Then the problem of integrating information from independent sources into
92 to one probability density $f_{\mathbf{I}}(\mathbf{m})$ is given by

$$f_{\mathbf{I}}(\mathbf{m}) \propto \prod_i^N f_{I_i}(\mathbf{m}). \quad (2)$$

93 In the present context \mathbf{m} reflects model parameters describing a reservoir
94 model, and I_1, I_2, \dots reflect different sources of information available (e.g. from
95 expert information, well log data, training image and geophysical data).

96 Here, the special case is considered where all information available refers
97 directly to the model parameters. The reason for this is two-fold: First,
98 most (any) geostatistical simulation algorithms allow, in principle, to take
99 such information into account as “soft” information (Mariethoz and Caers,
100 2014). Second, working with reservoir models, a lot of information about the

101 model parameters of interest can be available in form of direct measurements
 102 from well logs, inverted well logs parameters, or indirectly from geophysical
 103 data inverted into information about the model parameters \mathbf{m} . Barfod et al.
 104 (2016) present a recent example of how to do this, by establishing an at-
 105 las (applicable in Denmark) that can be used to translate resistivity values
 106 (found through inversion of airborne EM data) into lithological/hydrological
 107 units with associated uncertainty.

108 Three types of information are available in a typical MPS based geosta-
 109 tistical data integration problem:

110 I_{TI} *Information from a training image.* This can be information from out-
 111 crops, previous analysis, well log analysis, expert information which
 112 is quantified through a geostatistical model describing (spatial) co-
 113 dependence between model parameters.

114 I_{hard} *Hard data.* Direct observation of one or more model parameters, with-
 115 out any associated uncertainty.

116 I_{soft} *Soft data.* Direct observation of one or more model parameters, with
 117 an associated uncertainty.

118 In case the information has been obtained independently, such a geostatistical
 119 problem is equivalent to the problem of inferring information about $f_{\mathbf{I}}(\mathbf{m})$
 120 given by

$$f_{\mathbf{I}}(\mathbf{m}) \propto f_{I_{TI}}(\mathbf{m})f_{I_{hard}}(\mathbf{m})f_{I_{soft}}(\mathbf{m}). \quad (3)$$

121 Høyer et al. (2017) present one example of combining these three types of
 122 information into one stochastic model.

123 In principle there is no need to distinguish between hard and soft data, as
 124 both are simply data that provide information about the model parameters.
 125 So, a general geostatistical data integration problem can be formulated as

$$f_{\mathbf{I}}(\mathbf{m}) \propto f_{I_{TI}}(\mathbf{m})f_{I_{data}}(\mathbf{m}). \quad (4)$$

126 *Spatially independent 'data'.* For many geostatistical data integration prob-
 127 lems, the information about each model parameter is assumed spatially in-
 128 dependent, such that

$$f_{I_{data}}(\mathbf{m}) = \prod_{i=1}^M f_{I_{data}}(m_i). \quad (5)$$

129 From hereon, the term ‘soft information’ about the model parameters is de-
 130 fined through equation (5). The general data integration problem of equation
 131 (4) is then reduced to

$$f_{\mathbf{I}}(\mathbf{m}) \propto f_{TI}(\mathbf{m}) f_{data}(\mathbf{m}) = f_{TI}(\mathbf{m}) \prod_{i=1}^M f_{data}(m_i). \quad (6)$$

132 Equation (6) represent the probability distribution that most sequential sim-
 133 ulation based MPS methods suggest to sample from, by combining informa-
 134 tion from a geostatistical model with ‘hard’ (certain) and ‘soft’ (uncertain)
 135 data. From hereon different methods, existing and new, will be discussed
 136 that allow sampling from equation (6).

137 2.1. Markov chain Monte Carlo sampling of $f_{\mathbf{I}}(\mathbf{m}) \propto f_{TI}(\mathbf{m}) f_{data}(\mathbf{m})$

138 Sampling methods such as the extended Metropolis sampler provides a
 139 general, but computationally expensive, approach for sampling the product of
 140 two (or more) probability densities, both in form of equation (4) (accounting
 141 for spatially dependent information on the model parameters) and (6) (as-
 142 suming spatially independent information on the model parameters) Hansen
 143 et al. (2016a). Running the extended Metropolis algorithm consists of, in
 144 this case, sampling $f_{TI}(\mathbf{m})$ through a random walk, and accepting moving
 145 between proposed models based on acceptance criteria computed from the
 146 relative change in $f_{data}(\mathbf{m})$. Details on how to use the extended Metropolis
 147 sampler to sample from equation (4) and (6) can be found in e.g. Hansen
 148 et al. (2008, 2013, 2016a).

149 2.2. Sequential simulation of $f_{\mathbf{I}}(\mathbf{m}) \propto f_{TI}(\mathbf{m}) f_{data}(\mathbf{m})$

150 Sequential simulation (Alabert et al., 1989), also known as the conditional
 151 distribution method (Devroye, 1986), is commonly used in geostatistics to
 152 sample from $f_{I_{TI}}(\mathbf{m})$ and (conditional to data) $f_{\mathbf{I}}(\mathbf{m}) \propto f_{I_{TI}}(\mathbf{m}) f_{I_{data}}(\mathbf{m})$
 153 as in equation (6). In brief, sequential simulation consists of sequentially
 154 visiting and simulating all model parameters, possibly in random order. At
 155 the location of each model parameter m_i , the value of m_i is simulated (as
 156 m_i^*) conditional to all known information and all previously simulated model
 157 parameters, \mathbf{m}_c , as a realization from

$$f_{\mathbf{I}}(m_i | m_1^*, \dots, m_{i-1}^*) = f_{\mathbf{I}}(m_i | \mathbf{m}_c) = f_{TI}(m_i | \mathbf{m}_c) f_{data}(\mathbf{m}). \quad (7)$$

In case the available data are spatially independent, as in equation (6), the conditional distribution in equation (7) becomes

$$f_{\mathbf{I}}(m_i|\mathbf{m}_c) \approx f_{TI}(m_i|\mathbf{m}_c) \prod_{i=1}^M f_{data}(m_i) \quad (8)$$

Numerous methods based on sequential simulation has been developed in the geostatistical community that allow sampling from a wide variety of multiple-point statistical models inferred from a training image such as given by $f_{TI}(\mathbf{m})$ (Guardiano and Srivastava, 1993; Strebelle, 2000; Mariethoz et al., 2010; Straubhaar et al., 2011; Hansen et al., 2016b)) These methods differ in how the realization m_i^* of the conditional distribution in equation (7) is generated. Most of these methods allow, to some degree, to take into account direct information about the model parameters, hard and soft. In the following the ENESIM, SNESIM and Direct Sampling (DS) methods will be considered.

3. A synthetic example

In order to analyze the use of conditional information with sequential simulation algorithms based on MPS, a synthetic case study is designed. Figure 1a shows a training image (from Strebelle (2000), used to define $f_{TI}(\mathbf{m})$), consisting of pixels within (black) and outside (red) a channel structure, from which a reference model is generated as a realization in a 30x30 pixel grid, Figure 1b, using the ENESIM algorithm (Guardiano and Srivastava, 1993; Hansen et al., 2016b). The 25 closest previously simulated data are used to compute the conditional distribution at each step of the sequential simulation.

Simple smoothing of the reference realization in Figure 1b is performed in order to obtain an exhaustive map of 'soft' data that quantifies the local probability of each pixel belonging to a channel structure through $f_{I_{d1}}(\mathbf{m})$, Figure 2a. From this exhaustive set of soft data, a subset of 10 and 3 randomly chosen soft data points are considered as $f_{I_{d2}}(\mathbf{m})$ and $f_{I_{d3}}(\mathbf{m})$ and shown in Figures 2b-c.

The dense data set, I_{d1} , mimic an exhaustive set of information, as obtained from for example inversion of a densely sampled electromagnetic data set, as considered extensively by Barfod et al. (2016). The two sparse data sets, I_{d2} and I_{d3} , mimic information from well logs at different spatial density,

188 as considered by for example Høyer et al. (2017). Note that the two sparse
 189 sets of soft data, quantified by $f_{I_{d2}}(\mathbf{m})$ and $f_{I_{d3}}(\mathbf{m})$, can both be regarded as
 190 an exhaustive set of soft data with a uniform distribution everywhere a soft
 191 data is not explicitly defined.

192 In the following existing and new methods for sampling $f_{I_{TI}, I_{d1}}(\mathbf{m})$, $f_{I_{TI}, I_{d2}}(\mathbf{m})$,
 193 and $f_{I_{TI}, I_{d3}}(\mathbf{m})$, will be analyzed and compared.

194 [Figure 1 about here.]

195 [Figure 2 about here.]

196 4. A ‘reference’ solution - sampling from $f_{I_{TI}, I_d}(\mathbf{m}) = f_{I_{TI}}(\mathbf{m})f_{I_{data}}(\mathbf{m})$

197 The extended Metropolis algorithm is used to sample from $f_{\mathbf{I}}(\mathbf{m})$ con-
 198 sidering the three soft data sets defined above. This provides a reference
 199 solution (in form of a sample from $f(\mathbf{m}|I_{TI}, I_{data})$), to which other solutions
 200 can be compared. In practice, the ENESIM algorithm is used to gener-
 201 ate realizations from $f(\mathbf{m}|I_{TI})$ that are then accepted using the Metropolis
 202 acceptance criterion based on the soft data. In this way, 600 independent
 203 realizations have been obtained from $f_{I_{TI}, I_{d1}}(\mathbf{m})$, $f_{I_{TI}, I_{d2}}(\mathbf{m})$, and $f_{I_{TI}, I_{d3}}(\mathbf{m})$,
 204 using the SIPPI Matlab toolbox (Hansen et al., 2013). The corresponding
 205 probability of locating a channel, obtained using the above described algo-
 206 rithm are shown in Figure 3a-c. These results will be used as a reference for
 207 comparison.

208 [Figure 3 about here.]

209 5. Existing sequential simulation methods, using the Markov prop- 210 erty

Well known MPS algorithms such as ENESIM and SNESIM allow con-
 ditioning to uncertain data (Strebelle, 2000; Remy et al., 2008). In practice,
 most all MPS based sequential simulation algorithms use only co-located soft
 data (i.e. soft data located at the same position in space as the model pa-
 rameter m_i being simulated) when evaluating equation (8). The rest of the
 soft data are being ignored (see e.g. Strebelle (2000); Liu (2006); Remy et al.
 (2008)). In this case the marginal conditional probability being sampled
 during sequential simulation is reduced from equation (8) to

$$f_{\mathbf{I}}(m_i|\mathbf{m}_c) \propto f_{TI}(m_i|\mathbf{m}_c) f_{data}(m_i) \quad (9)$$

211 This assumption is similar to the Markov property assumed for sequential
 212 Gaussian co-simulation, as proposed by Almeida and Journel (1994). There-
 213 fore the approximation in equation (9) is referred to as using a Markov prop-
 214 erty to handle the soft data. Equation ((9)) assumes that the source of
 215 the information from the training image, $f_{TI}(m_i|\mathbf{m}_c)$, and the ‘soft’ data,
 216 $f_{data}(m_i)$, are independent. If this is not the case, one can use e.g. the tau-
 217 model to explicitly model the dependence between the two types of available
 218 information (Journel, 2002; Krishnan, 2008). The amount of dependency is
 219 controlled by the tau factor. Estimation of a proper value of the tau factor,
 220 can in itself be a challenging task, and is not considered further here.

221 The complexity related to implementing an algorithm that samples from
 222 equation (9) depends on the choice of MPS algorithm. Below we briefly
 223 describe these differences for a number of widely used methods. We refer to
 224 Mariethoz and Caers (2014) for a general description of MPS algorithms.

225 5.1. ENESIM and the Markov property

226 Using ENESIM the full conditional distribution $f_{TI}(m_i|\mathbf{m}_c)$ is explicitly
 227 computed at each iteration by scanning the whole training image. Therefore
 228 evaluation of equation (9) is straightforward to implement.

229 5.2. SNESIM type algorithms and the Markov property

230 SNESIM (Strebelle, 2000), and related IMPALA (Straubhaar et al., 2011),
 231 type simulation algorithms scans the training image only once, for a number
 232 of predefined sets of conditional point patterns. The frequency of occurrence
 233 for each pattern is then stored in memory. At each iteration in the sequential
 234 simulation $f_{TI}(m_i|\mathbf{m}_c)$ is then obtained from memory, and hence evaluation
 235 of equation (9) straightforward.

236 However, SNESIM also makes use of so-called multiple-grids, that is
 237 needed to allow reproducing correlations over long distances, while at the
 238 same time reducing the memory requirements (Tran, 1994). This introduces
 239 a challenge when conditioning hard and soft data are available, as condi-
 240 tional data may not be available on a specific coarse grid being simulated.
 241 To remedy this, so-called re-location of hard data has been suggested. When
 242 simulating on a coarse grid, the closest hard data at finer grids are re-located
 243 to the coarse simulation grid as a hard data. Then conditional simulation
 244 is performed in the coarse grid. Finally after, simulation of the coarse grid
 245 the hard data values at the notes of the re-located data, are removed, and
 246 set as un-sampled. See details in Strebelle (2000); Remy et al. (2008). Here,

re-location of the soft data has been implemented in the SNESIM implementation in MPSlib (Hansen et al., 2016b), in a manner similar to the approach used for hard data. Note that in case the uncertain/soft data are exhaustively available, no relocation is needed. Straubhaar and Malinverni (2014) propose an alternative approach for handling conditional data with multiple grids, that can lead to less artifacts.

5.3. Handling co-located soft data using DS

Using the DS algorithm $f_{TI}(m_i|\mathbf{m}_c)$ is never explicitly computed, instead a realization from $f_{TI}(m_i|\mathbf{m}_c)$ is obtained directly from the training image. This means the DS algorithm cannot take co-located soft data into account simply by evaluating equation (9).

Biver et al. (2014) and Straubhaar et al. (2016) suggest an approach that aims to reproduce the local proportions within a data neighborhood, as provided by I_{soft} (for data of both point and volume support). In their approach uncertainty of the soft data is not taken into account explicitly as defined in equation (9). Instead we propose to use a simple application of the extended rejection sampler that allows the direct sampling algorithm to generate a realization of $f_{\mathbf{I}}(m_i|\mathbf{m}_c) = f_{TI}(m_i|\mathbf{m}_c) f_{data}(m_i)$, using the exact same conditions as ENESIM and SNESIM. Numerical implementation consists of replacing the step of scanning the training image for the first matching conditional data event \mathbf{m}_c , with the following algorithm

- Start loop
 1. Obtain a realization, m_i^* , of $f_{TI}(m_i|\mathbf{m}_c)$ (by scanning the training image).
 2. Accept m_i^* as a realization of $f_{TI}(m_i|\mathbf{m}_c)f_{data}(m_i)$ with probability $P_{acc} = \frac{f_{data}(m_i=m_i^*)}{\max(f_{data}(m_i))}$.
- End loop (when m_i^* is accepted) .

$\max(f_{data}(m_i))$ is the maximum probability of any possible value of m_i . This will ensure that m_i^* will be a realization of $f(m_i|\mathbf{m}_c)f_{data}(m_i)$ as given in equation (9). This rejection step has been implemented in the GENESIM algorithm in MPSlib (Hansen et al., 2016b), which is a generalized implementation of the ENESIM algorithm, in which the conditional distribution is based on any number N_c of observed matches. If $N_c = 1$, the GENESIM algorithm will in practice perform similar to the DS algorithm (Hansen et al.,

281 2016b). In the remainder, when we refer to the DS algorithm we use the
282 GENESIM algorithm with $N_c = 1$.

283 5.4. *Conditional ENESIM/SNESIM/DS simulation using the Markov prop-* 284 *erty*

285 Using the ENESIM algorithm and the Markov property for conditioning
286 to 'soft' data, 600 independent realizations are generated and the correspond-
287 ing probability of locating a channel computed. The results are shown in
288 Figure 4 in case using a 'unilateral' (i.e., raster scan) path (top, a)-c)), and
289 in case using a random path (bottom, d)-f)). Similar results obtained using
290 SNESIM are shown in Figure 5. No results are shown using DS as they are
291 essentially identical to those obtained using ENESIM in Figure 4.

292 [Figure 4 about here.]

293 [Figure 5 about here.]

294 Figure 4 reveals that the simulation results lack information as compared
295 to the full solution (Figure 3). This is most severe in case soft data are sparse
296 in which case little to no information from the soft data seems to have been
297 taken into account (Figure 4b-c and 4e-f). So, while it is rather straight-
298 forward to account for uncertain information about the model parameters
299 using the Markov property (as also stated by Straubhaar et al., 2016), it
300 may not be a viable approach using either a sequential or random simulation
301 path. Below we propose two alternative approaches to better account for the
302 available uncertain/soft data.

303 6. Suggestion 1: preferential simulation path

It has long been known that the choice of simulation path affects the
realizations generated using sequential simulation (Strebelle, 2000; Liu and
Journel, 2004; Daly, 2005; Mariethoz and Renard, 2010; Daly, 2005). One
problem of using either the unilateral or random path with the Markov prop-
erty as considered above, is that information from highly informed model pa-
rameters located very close to a model parameter, for which the conditional
distribution is computed, is disregarded. Consider two direct observations
 $f(m_i = 1) = 0.999$ and $f(m_j = 1) = 1$ (which implies $f(m_i = 0) = 0.001$ and

$f(m_j = 0) = 0$, as the training image only allows $k=2$ possible outcomes). The entropy

$$E(f(m)) = - \sum_k f(m = m^k) \log_2(f(m = m^k)), \quad (10)$$

is a measure of uncertainty of the information provided by $f(m)$ (Reza, 1961). With $k=2$ possible outcomes, the maximum entropy is given by $E_{max}(f(m)) = 1$. A base of 2 is used for the logarithm in equation (10), which is a natural choice with $k=2$ possible outcomes. A base of k , would be a natural choice for a training images with k possible outcomes. A simple measure of the ‘certainty’ of the information provided by $f(m)$ can then be formulated as

$$C(f(m)) = 1 - \frac{E(f(m))}{E_{max}} \quad (11)$$

This leads to $C(f(m_i)) = 0.99$ and $C(f(m_j)) = 1$. Thus, these two types of information provide almost the same information. However, in a typical implementation of an MPS algorithm (as discussed above) $f(m_j = 1) = 1$ is treated as hard data, and the value of m_j is fixed at $m_j^* = 1$ prior to simulation. This means that $m_j^* = 1$ will be used as conditional data in any subsequent step of the sequential simulation algorithm.

The information provided by $f(m_i = 1) = 0.99$ will however be treated as uncertain/soft data, and will (using the Markov property) only come into use when the simulation algorithm visits m_i , when a realization of $f(m_i|\mathbf{m}_c)$ has to be generated. Depending on the choice of random path this can happen early or late in the simulation process. If it happens early, then the information in $f(m_i = 1) = 0.99$ will affect the simulated value of relatively many model parameters. If it happens late in the process the information will only affect relatively few model parameters. Due to the use of the Markov property, the amount of information used for a given model parameter is closely related to the choice of random path. This is the reason for the relatively poor conditioning to the soft data obtained using sequential simulation with the Markov Property, using both a unilateral and random path as seen in Figures 4-5.

To remedy some of these problems the use of a *preferential* random path is suggested, where model parameters with soft data with high information content is visited preferentially to soft data with lower information content.

In practice the preferential path can be computed prior to running the sequential simulation algorithm. First, the entropy $E(f_{data}(m_i))$ is computed

for all soft data. Then, a pseudo random path is given by ordering all the model parameters in ascending order by $order_i$ given by

$$order_i = r_i - 1 + I_{fac} C(f(m)), \quad (12)$$

where r_i is a random number between 0 and 1. I_{fac} is a factor that controls the 'randomness' associated to the information content. If $I_{fac} = 0$ all model parameters with soft data are visited at random (in no specific order), before model parameters with no soft data are visited. When I_{fac} is high then locations with soft data are visited in order of decreasing information content. In the following $I_{fac} = 4$ is used.

6.1. Conditional ENESIM/SNESIM/DS simulation using the Markov property and the preferential path

Figures 6, 7, and 8 show the probability of locating a channel conditional to the three data sets, based on 600 realizations generated by ENESIM, DS, and SNESIM using a preferential path. If $P_{mcmc}(channel)$ and $P(channel)$ refer to the posterior probability of locating a channel in each pixel using the reference MCMC approach and a specific choice of simulation, then Tables 3-1 summarize the relative difference in L2-norm as $L_2(P_{mcmc}(channel) - P(channel))/L_2(P_{mcmc}(channel))$, for different simulation choices and choice of simulation algorithm. A number close to 0 suggests that simulation results (in form of the posterior probability of locating a channel) is very close to the results obtained using the reference MCMC approach, Figure 3, while a higher number will refer to less similarity. From hereon we refer to this quantity as the 'relative L2 norm'.

6.1.1. ENESIM

Using ENESIM with a preferential path conditional to I_{d1} , it is clear that not as much information is extracted from the uncertain data, Figure 6a, as is the case using full Monte Carlo sampling, Figure 3a. This difference is due the fact that the Markov property is not used as part of the Monte Carlo sampling, which will lead to better resolved channel structures. However, significantly more information is extracted than when using an unilateral or random simulation path, see Figures 4a and 4d. Table 1 also shows a significant drop in the relative L2-norm using the preferential path (0.43 vs 0.69 using a random path).

In the case of sparse data (I_{d2} and I_{d3}) the use of a preferential path provides results, Figure 6b-c, that are close to indistinguishable from the full

369 non-Markov solution, obtained using Monte Carlo sampling, Figure 3b-c.,
370 with a corresponding small L2 norm, Table 1.

371 [Figure 6 about here.]

372 6.1.2. DS

373 The results obtained using DS, Figure 7, are similar to the results ob-
374 tained using the ENESIM algorithm, Figure 6, and quantified in Table 2,
375 illustrating that the use of the rejection sampler with DS works as intended.

376 [Figure 7 about here.]

377 6.1.3. SNESIM

378 Comparing Figure 5 to 8 it is evident that the use of a soft data relocation
379 and a preferential path with SNESIM allow much better reproduction of
380 uncertain data. However, some effects of using multiple grids and re-location
381 persist, which is the reason of the relative high relative L2 norm of 0.23 using
382 SNESIM compared to 0.09 using ENESIM and DS in case conditioning to
383 I_{d3} , see Tables 1-3.

384 One simple approach to remedy some of the effects of re-location of soft
385 (and hard) data, is to make use of ENESIM type algorithms to perform the
386 simulation on coarser grids, as suggested by Strebelle (2000) to avoid prob-
387 lems related to hard-data relocation. Another approach could be to consider
388 applying the approach proposed by Straubhaar and Malinverni (2014) also
389 to soft/uncertain data, to avoid artifacts caused by the use of multiple grids.

390 [Figure 8 about here.]

391 Tables 1-3 highlights that in general the use of the preferential path, with
392 the Markov assumption considering only colocated data, significantly reduces
393 the relative L2 norm. Further Tables 1-3 suggest the difference in simulation
394 time using the preferential path compared to using the random path is small.

395 The preferential path emulates what has been done in practice since the
396 first simulation algorithms were developed. If ‘hard’ information is available,
397 i.e. certain information about the model parameters, then these model pa-
398 rameters will be visited before other model parameters using the preferential
399 path. This is equivalent to simply assigning the hard data to the correspond-
400 ing model parameters prior to starting the simulation. It is also related to

401 simulating model parameters with soft information prior to other data, as
 402 proposed by Soares et al. (2016) in case using Gaussian direct sequential
 403 simulation.

404 Liu and Journel (2004) also suggest to choose the random path guided
 405 by the information content. Unlike the present work, where the path is
 406 guided by the information content of the soft data, they suggest to guide the
 407 path based on the conditional information from the training image, i.e. from
 408 $f_{TI}(m_i|\mathbf{m}_c)$. They demonstrate that such a path better reproduces large
 409 scale connected structures, compared to using a random simulation path.

410 7. Suggestion 2: Avoiding the Markov property

411 The Markov property can in principle be avoided entirely, to allow con-
 412 sidering more than just co-located soft information, while still using the se-
 413 quential simulation approach, and using a fully random path.

414 7.1. DS conditional to non-located soft data

415 The extended rejection sampler used above to allow the DS algorithm to
 416 condition to co-located uncertain/soft data, can be generalized to account
 417 for, in principle, all soft data, without the need for the Markov property. A
 418 sample of equation (6) can be obtained at each iteration of the sequential
 419 simulation algorithm using the extended rejection sampler as follows:

- 420 • Start loop
 - 421 1. Obtain a realization, m_i^* , of $f_{TI}(m_i|\mathbf{m}_c)$.
 2. Accept m_i^* as a realization of $f_{TI}(m_i|\mathbf{m}_c) \prod_{i=1}^M f_{data}(m_i)$ with prob-
 ability

$$P_{acc} = \frac{\prod_{i=1}^{N_s} f_{data}(m_i = m_i^*)}{\prod_{i=1}^{N_s} \max(f_{data}(m_i))} \quad (13)$$

- 422 • Continue loop (until m_i^* is accepted).

423 N_s refers to the closest N_s soft/uncertain data. In case $N_s = \infty$, the above
 424 will sample from full probability density given in equation (6), without the
 425 Markov assumption. Hence, results should be comparable to using the Monte
 426 Carlo based sampling approach.

427 In practice, due to both CPU requirements and the limited size of the
 428 training image, N_s can be chosen to use limited set of conditional soft data,
 429 while providing simulation results similar to using a full neighborhood, using
 430 much less computational power.

431 *Conditional simulation to soft data, without the preferential path.* When con-
432 ditioning to non-colocated soft data, the use of the preferential path should,
433 in principle, no longer be needed in order to condition to soft/uncertain data.
434 Figure 9 shows the probability of locating a channel in case using a random
435 path, and the 3 closest soft/uncertain data using DS type simulation using
436 the rejection sampling approach described above.

437 In general the resolution is better than using a unilateral or random
438 with the Markov property, but worse than using a preferential path and
439 the Markov property (see e.g. Table 2)

440 This is due to conditioning to soft data becoming more difficult if a lot
441 of model parameters are visited, and hence simulated, prior to visiting the
442 location of the soft data. In this case the 'hard' simulated data will take
443 precedence over the soft data, unless a non-perfect match to the hard data
444 is allowed. This is one reason why the use of the preferential path may be
445 useful even when conditioning to non-colocated soft data.

446 *Conditional simulation to soft data, with the preferential path..* Another rea-
447 son to use the preferential path in this case is that it can lead to a com-
448 putationally more efficient simulation algorithm. Using a random path, one
449 will have to evaluate the rejection sampler described above, at all iterations
450 until all model parameters with soft data have been simulated. If using a
451 preferential random path, one need only evaluate the rejection sampling step
452 above, until all soft data has been evaluated. Thus, only for the first 3 and
453 11 iterations considering I_{d3} and I_{d2} .

454 Figure 10 shows results obtained running the DS algorithm to generate
455 600 independent realizations, using $N_s = 1$ (top), $N_s = 3$ (middle), and
456 $N_s = 11$ without the Markov property, with a preferential path. Table 2
457 shows the corresponding relative L2-norm and simulation time.

458 For the most sparse data set, I_{d3} , a subtle difference can be identified
459 comparing figure 10c) ($N_s = 1$) and 10f) ($N_s = 3$), leading to a slightly
460 smaller relative L2 norm. Considering $N_s = 3$, the probability of locating a
461 channel is slightly larger than when using $N_s = 1$. In general, there is little
462 to no difference using $N_s = 3$ or $N_s = 11$ conditioning to sparse soft data,
463 I_{d2} and I_{d3} .

464 It is also clear that when conditioning to the exhaustive soft data set,
465 I_{d1} , the amount of information extracted from the soft data (as quantified in
466 Table 2), increases as the number of conditioning soft data increases, Figure
467 10a,d,g. For this conditional data set, the best result (i.e. that best resemble

468 the reference solution) is obtained using 11 conditional data, Figure 10g.

469 This algorithm, as any rejection algorithm, will only be feasible if the
470 number of conditioning soft data is small. Alternatively one can make use
471 of only a limited number of the closest soft data, to allow a better use of the
472 soft data, while limiting the computational needs.

473 Note in Table 2 that when using a random simulation path, and non-
474 colocated soft data, results in a significant increase in simulation times (a
475 factor of 1-8) when using more non-located soft data as compared to only
476 one soft data. Using the preferential path the simulation times is only a few
477 percent larger using 25 conditional soft data, as opposed to 1 conditional soft
478 data, in the case of conditioning to e.g I_{d2} .

479 7.2. ENESIM/GENESIM conditional to non-colocated soft data

480 The ENESIM/GENESIM algorithm can be also generalized to sample
481 conditionally to non-colocated soft data. In this case the whole (using EN-
482 ESIM) or a limited random part (using GENESIM) of the training image
483 is scanned at each iteration. For each match of a hard data, the specific
484 value of the centered node in the training image, is associated with the (soft)
485 probability $\prod_{i=1}^{N_s} (f_{data}(m_{i|j}))$. j is the position in the training image and
486 $m_{i|j}$ refer to the value of the location of the soft data relative to the current
487 location in the training image. Conditioning to non-colocated soft data as
488 described here have been implemented in the MPSlib codes in the GENESIM
489 algorithm, Hansen et al. (2016b).

490 Simulation times and relative L2 norms using GENESIM type simula-
491 tion are, for reference, presented in Table 1, for the same conditional data
492 sets considered by DS in Table 2. Even though the handling of soft data
493 in DS and ENESIM type simulation is quite different, the main difference
494 between the two algorithms are with respect to simulation times, which is
495 expected. The GENESIM algorithm can be used to scan only a limited set
496 conditional evenets, which is much faster than using ENESIM that scans the
497 entire training image at each iteration.

498 7.3. SNESIM conditional to non-colocated soft data

499 While SNESIM can in principle also be generalized to account for non-
500 colocated soft data, problems related to re-location persist, and search times
501 scanning the search tree will become large. Therefore, we do not pursue this
502 approach further, and leave this for potential future research.

503 [Figure 9 about here.]

504 [Figure 10 about here.]

505 8. Conclusion

506 MPS based sequential simulation algorithms allow for a computationally
507 efficient approach to the problem of integration of probabilistic information
508 from different sources. However, the traditionally used Markov property,
509 using only co-located uncertain soft data, leads to realizations that do not
510 fully take into account the information of the soft data. The problem is
511 most severe when sequential simulation is performed with soft information
512 available at sparse locations. Two methods have been proposed that allow
513 taking soft data properly into account.

514 First, a simulation path preferential to 1D marginal entropy/information
515 content of soft data has been proposed. This allows much better handling of
516 especially scattered soft data. The preferential path is trivial to use with the
517 ENESIM algorithm. Using a simple rejection step to account for soft data,
518 it can be easily implemented with the DS algorithm. It is straightforward
519 to use with the SNESIM algorithm, but re-location of soft data is suggested
520 due to the use of multiple grids.

521 Second, an approach is suggested that avoid the Markov-property, such
522 that non co-located soft data can be considered, that can be used with any
523 of the ENESIM and DS algorithms. Combined with using a preferential path
524 this leads to a conditional simulation algorithm that properly conditions to
525 the soft data, while at the same time being computationally much more
526 viable than using McMC sampling methods.

527 9. Acknowledgments

528 This research has been funded by two projects (113-2013-1,53-2014-3)
529 funded partly by the Danish High Technology Foundation. MPS simulation
530 codes and examples are available at <https://github.com/ergosimulation/mpslib/>.

531 10. References

532 Alabert, F., et al., 1989. Non-gaussian data expansion in the earth sciences.
533 Terra Nova 1 (2), 123–134.

534 Almeida, J. S., Journel, A. G., 1994. Joint simulation of multiple variables
535 with a markov-type coregionalization model. *Mathematical Geology* 26 (5),
536 465–588.

537 Barfod, A., Mller, I., Christiansen, A., 2016. Compiling a national resistivity
538 atlas of denmark based on airborne and ground-based transient electro-
539 magnetic data. *Journal of Applied Geophysics*.

540 Biver, P., Mariethoz, G., haar, J., Chugunova, T., Renard, P., 2014. Handling
541 soft probabilities in multiple point statistics simulation. In: *Mathematics*
542 *of Planet Earth*. Springer, pp. 69–72.

543 Cordua, K. S., Hansen, T. M., Mosegaard, K., 2012. Monte Carlo full wave-
544 form inversion of crosshole GPR data using multiple-point geostatistical a
545 priori information. *Geophysics* 77, H19–H31.

546 Daly, C., 2005. Higher order models using entropy, markov random fields and
547 sequential simulation. *geostatistics Banff 2004*, 215–224.

548 Deutsch, C., Journel, A., 1992. *GSLIB: Geostatistical Software Library and*
549 *User’s Guide*. Oxford University Press.

550 Devroye, L., 1986. Sample-based non-uniform random variate generation.
551 In: *Proceedings of the 18th conference on Winter simulation*. ACM, pp.
552 260–265.

553 Guardiano, F. B., Srivastava, R. M., 1993. Multivariate geostatistics: beyond
554 bivariate moments. In: *Geostatistics Troia 92*. Springer, pp. 133–144.

555 Hansen, T., Cordua, K., Looms, M., Mosegaard, K., 2013. SIPPI: a Matlab
556 toolbox for sampling the solution to inverse problems with complex prior
557 information: Part 1, methodology. *Computers & Geosciences* 52, 470–480.

558 Hansen, T., Cordua, K., Zunino, A., Mosegaard, K., 2016a. Probabilistic in-
559 tegration of geo-information. In: Moorekamp, M. (Ed.), *Integrated Imag-*
560 *ing of the Earth: Theory and Applications*. AGU, pp. 93–116.

561 Hansen, T. M., Cordua, K. C., Mosegaard, K., 2012. Inverse problems with
562 non-trivial priors - efficient solution through sequential Gibbs sampling.
563 *Computational Geosciences* 16 (3), 593–611.

564 Hansen, T. M., Mosegaard, K., Cordua, K. C., 2008. Using geostatistics
565 to describe complex a priori information for inverse problems. In: Ortiz,
566 J. M., Emery, X. (Eds.), VIII International Geostatistics Congress. Vol. 1.
567 Mining Engineering Department, University of Chile, pp. 329–338.

568 Hansen, T. M., Vu, L. T., Bach, T., 2016b. MPSLIB: A C++ class for
569 sequential simulation of multiple-point statistical models. *Software X* 5,
570 127–133.

571 He, X., Sonnenborg, T., Jørgensen, F., Jensen, K. H., 2014. The effect of
572 training image and secondary data integration with multiple-point geo-
573 statistics in groundwater modelling. *Hydrology and Earth System Sciences*
574 18 (8), 2943–2954.

575 Høyer, A.-S., Jørgensen, F., Foged, N., He, X., Christiansen, A., 2015. Three-
576 dimensional geological modelling of aem resistivity dataa comparison of
577 three methods. *Journal of Applied Geophysics* 115, 65–78.

578 Høyer, A.-s., Vignoli, G., Hansen, T. M., Keefer, D. A., Jørgensen, F., et al.,
579 2017. Multiple-point statistical simulation for hydrogeological models: 3-
580 D training image development and conditioning strategies. *Hydrology and*
581 *Earth System Sciences Discussions* 21, 6069–6089.

582 Irving, J., Singha, K., 2010. Stochastic inversion of tracer test and electrical
583 geophysical data to estimate hydraulic conductivities. *Water Resources*
584 *Research* 46.

585 Jaynes, E. T., 1968. Prior probabilities. *IEEE Transactions on systems sci-*
586 *ence and cybernetics* 4 (3), 227–241.

587 Jørgensen, F., Høyer, A.-S., Sandersen, P. B., He, X., Foged, N., 2015. Com-
588 bining 3D geological modelling techniques to address variations in geology,
589 data type and density—an example from southern denmark. *Computers &*
590 *Geosciences* 81, 53–63.

591 Journel, A., 1986. Constrained interpolation and qualitative information -
592 the soft kriging approach. *Mathematical Geology* 18 (3), 269–286.

593 Journel, A., 2002. Combining knowledge from diverse sources: An alternative
594 to traditional data independence hypotheses. *Mathematical geology* 34 (5),
595 573–596.

596 Koch, J., He, X., Jensen, K. H., Refsgaard, J. C., 2014. Challenges in con-
597 ditioning a stochastic geological model of a heterogeneous glacial aquifer
598 to a comprehensive soft data set. *Hydrology and Earth System Sciences*
599 18 (8), 2907–2923.

600 Krishnan, S., 2008. The tau model for data redundancy and information
601 combination in earth sciences: Theory and application. *Mathematical Geo-*
602 *sciences* 40 (6), 705.

603 Liu, Y., 2006. Using the snesim program for multiple-point statistical simu-
604 lation. *Computers & Geosciences* 32 (10), 1544–1563.

605 Liu, Y., Journel, A., 2004. Improving sequential simulation with a structured
606 path guided by information content. *Mathematical Geology* 36 (8), 945–
607 964.

608 Mariethoz, G., Kelly, B. F., 2011. Modeling complex geological structures
609 with elementary training images and transform-invariant distances. *Water*
610 *Resources Research* 47 (7).

611 Mariethoz, G., Renard, P., 2010. Reconstruction of incomplete data sets or
612 images using direct sampling. *Mathematical Geosciences* 42 (3), 245–268.

613 Mariethoz, G., Renard, P., Straubhaar, J., 2010. The direct sampling method
614 to perform multiple-point geostatistical simulations. *Water Resources Re-*
615 *search* 46 (11).

616 Mariethoz, P., Caers, P., 2014. *Multiple-point Geostatistics: Stochastic Mod-*
617 *eling with Training Images*. Wiley.

618 Mosegaard, K., Tarantola, A., 1995. Monte carlo sampling of solutions to
619 inverse problems. *J. geophys. Res* 100 (B7), 12431–12447.

620 Mosegaard, K., Tarantola, A., 2002. Probabilistic approach to inverse prob-
621 lems. *International Geophysics* 81, 237–265.

622 Remy, N., Boucher, A., Wu, J., 2008. *Applied Geostatistics with SGeMS: A*
623 *User’s Guide*. Cambridge University Press.

624 Renard, P., Straubhaar, J., Caers, J., Mariethoz, G., 2011. Conditioning fa-
625 cies simulations with connectivity data. *Mathematical Geosciences* 43 (8),
626 879–903.

- 627 Reza, F. M., 1961. An introduction to information theory. Courier Corpora-
628 tion.
- 629 Soares, A., Nunes, R., Azevedo, L., 2016. Integration of uncertain data in
630 geostatistical modelling. *Mathematical Geosciences*, 1–21.
- 631 Straubhaar, J., Malinverni, D., 2014. Addressing conditioning data in
632 multiple-point statistics simulation algorithms based on a multiple grid
633 approach. *Mathematical Geosciences* 46 (2), 187–204.
- 634 Straubhaar, J., Renard, P., Mariethoz, G., 2016. Conditioning multiple-point
635 statistics simulations to block data. *Spatial Statistics* 16, 53–71.
- 636 Straubhaar, J., Renard, P., Mariethoz, G., Froidevaux, R., Besson, O., 2011.
637 An improved parallel multiple-point algorithm using a list approach. *Math-*
638 *ematical Geosciences* 43 (3), 305–328.
- 639 Strebelle, S., 2000. Sequential simulation drawing structures from training
640 images. Ph.D. thesis, Stanford University.
- 641 Tahmasebi, P., Sahimi, M., Caers, J., 2014. Ms-ccsim: accelerating pattern-
642 based geostatistical simulation of categorical variables using a multi-scale
643 search in fourier space. *Computers & Geosciences* 67, 75–88.
- 644 Tarantola, A., 2005. Inverse problem theory and methods for model param-
645 eter estimation. SIAM.
- 646 Tarantola, A., Valette, B., 1982. Inverse problems= quest for information. *J.*
647 *geophys* 50 (3), 150–170.
- 648 Toftaker, H., Tjelmeland, H., 2013. Construction of binary multi-grid markov
649 random field prior models from training images. *Mathematical Geosciences*
650 45 (4), 383–409.
- 651 Tran, T. T., 1994. Improving variogram reproduction on dense simulation
652 grids. *Computers & Geosciences* 20 (7-8), 1161–1168.
- 653 Zinn, B., Harvey, C. F., 2003. When good statistical models of aquifer het-
654 erogeneity go bad: A comparison of flow, dispersion, and mass transfer in
655 connected and multivariate gaussian hydraulic conductivity fields. *Water*
656 *Resources Research* 39 (3).

657

[Table 1 about here.]

658

[Table 2 about here.]

659

[Table 3 about here.]

660 List of Figures

661	1	a) Training image. b) Reference realization. Pixel color refer	
662		to inside (black) and outside (red) a channel.	25
663	2	Soft data. a) Exhaustive, 900 soft data $f_{I_{d1}}(\mathbf{m})$, b) 10 soft	
664		data, $f_{I_{d2}}(\mathbf{m})$, and c) 3 soft data, $f_{I_{d3}}(\mathbf{m})$	26
665	3	Posterior probability of locating a channel, $P(m_i = 1 I_{TI}, I_d)$,	
666		obtained using the extended Metropolis sampler, conditional	
667		to the three sets of soft data a) Exhaustive, $d1$, b) 10 random	
668		soft data, $d2$, and c) 3 random soft data, $d3$	27
669	4	Posterior probability of locating a channel using the ENESIM	
670		algorithm with a top) unilateral and bottom) random path,	
671		conditional to a),d) $d1$, b),e) $d2$, and c),f) $d3$. Compare to	
672		Figure 3.	28
673	5	Posterior probability of locating a channel using the SNESIM	
674		algorithm with a top) unilateral, and bottom) random path,	
675		conditional to a),d) $d1$, b),e) $d2$, and c),f) $d3$	29
676	6	Posterior probability of locating a channel using the ENESIM	
677		algorithm with a preferential path, conditional to a) $d1$, b)	
678		$d2$, and c) $d3$. Compare to Figure 4 and the 'full' solution in	
679		Figure 3.	30
680	7	Posterior probability of locating a channel using the DS algo-	
681		gorithm with a preferential path, conditional to a) $d1$, b) $d2$, and	
682		c) $d3$	31
683	8	Posterior probability of locating a channel using the SNESIM	
684		algorithm with a preferential path, conditional to a) $d1$, b) $d2$,	
685		and c) $d3$	32
686	9	Posterior probability of locating a channel using the DS algo-	
687		gorithm using the 3 closest 'soft' data and the 25 closest previ-	
688		ously simulated data, with a random path, conditional to a)	
689		$d1$, b) $d2$, and c) $d3$	33
690	10	Posterior probability of locating a channel using the DS algo-	
691		gorithm using the closest 1 (a,b,c), 3 (d,e,f), and 11 (g,h,i)	
692		'soft'/uncertain data and the 25 closest previously simulated	
693		data, with a preferential path, conditional to $d1$ (a,d,f), $d2$	
694		(b,e,g), and $d3$. (c,f,h)	34

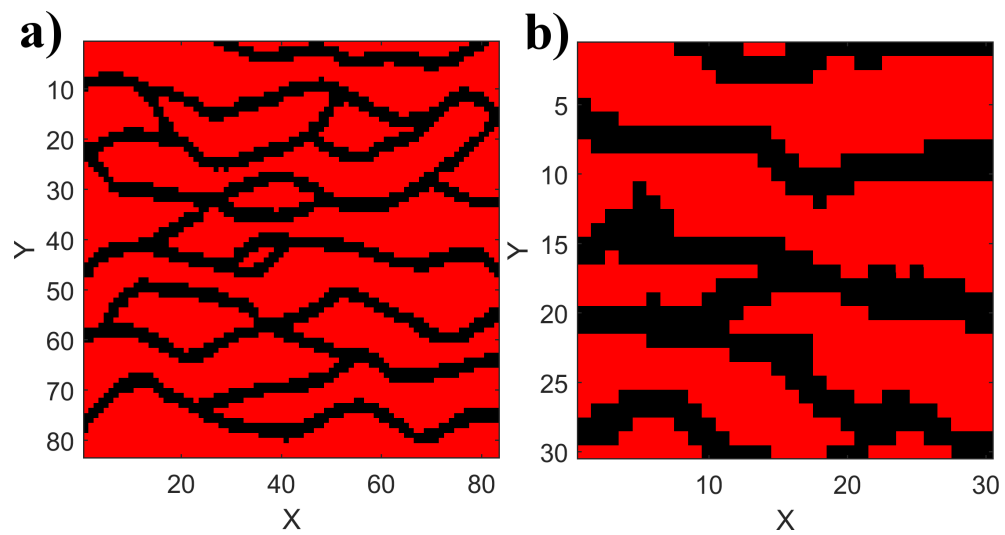


Figure 1: a) Training image. b) Reference realization. Pixel color refer to inside (black) and outside (red) a channel.

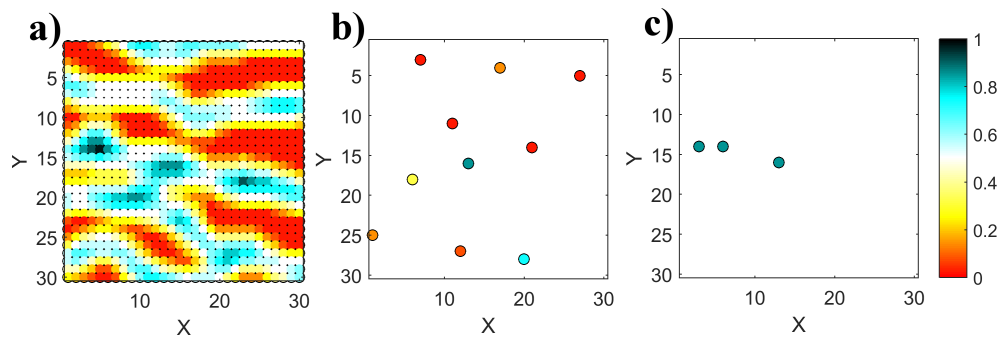


Figure 2: Soft data. a) Exhaustive, 900 soft data $f_{I_{d1}}(\mathbf{m})$, b) 10 soft data, $f_{I_{d2}}(\mathbf{m})$, and c) 3 soft data, $f_{I_{d3}}(\mathbf{m})$

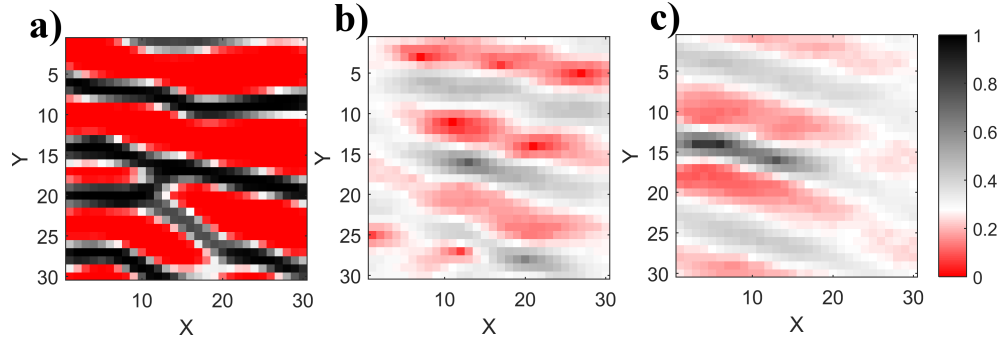


Figure 3: Posterior probability of locating a channel, $P(m_i = 1|I_{TI}, I_d)$, obtained using the extended Metropolis sampler, conditional to the three sets of soft data a) Exhaustive, $d1$, b) 10 random soft data, $d2$, and c) 3 random soft data, $d3$.

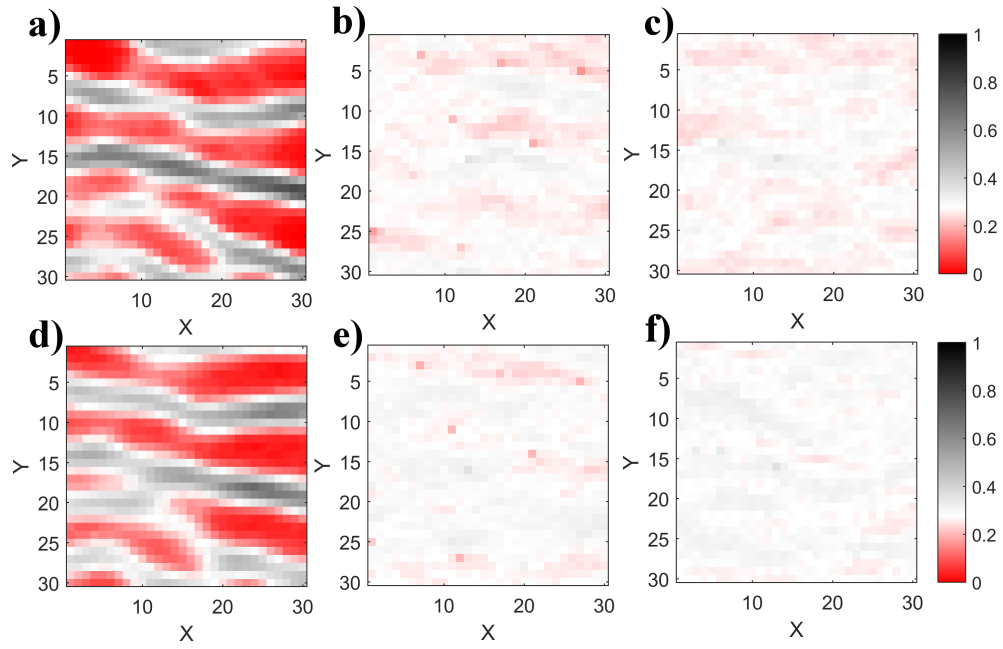


Figure 4: Posterior probability of locating a channel using the ENESIM algorithm with a top) unilateral and bottom) random path, conditional to **a)**,**d)** $d1$, **b)**,**e)** $d2$, and **c)**,**f)** $d3$. Compare to Figure 3.

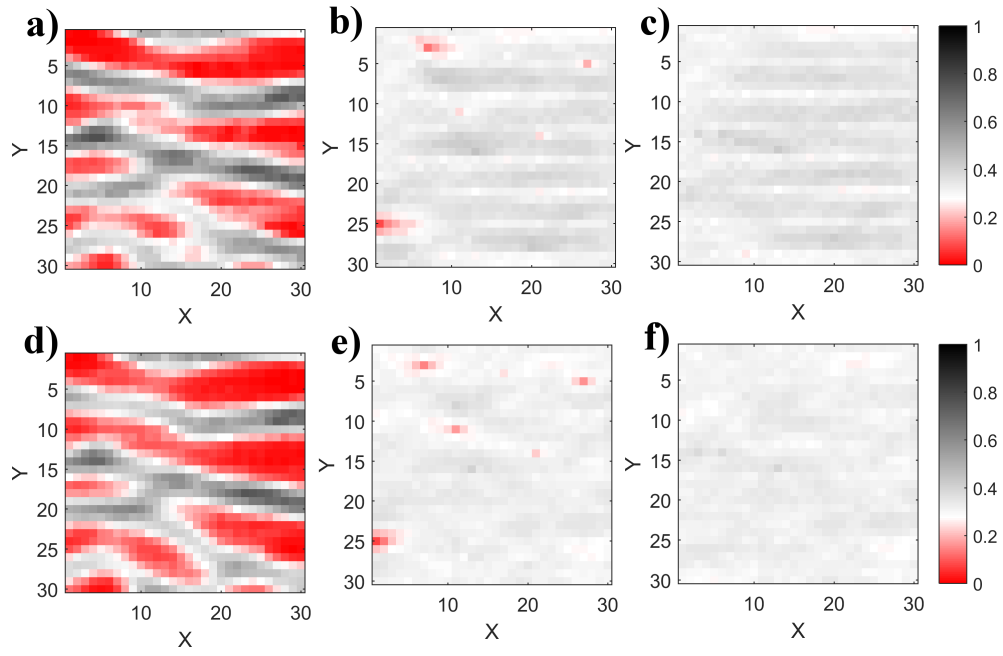


Figure 5: Posterior probability of locating a channel using the SNESIM algorithm with a top) unilateral, and bottom) random path, conditional to a),d) $d1$, b),e) $d2$, and c),f) $d3$.

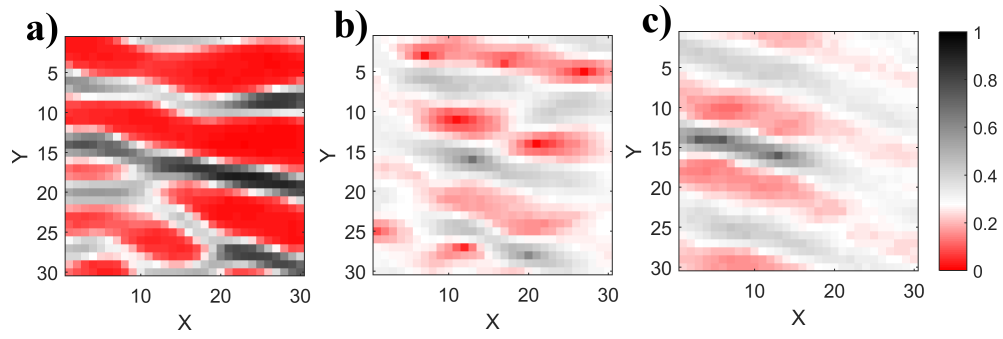


Figure 6: Posterior probability of locating a channel using the ENESIM algorithm with a preferential path, conditional to a) $d1$, b) $d2$, and c) $d3$. Compare to Figure 4 and the 'full' solution in Figure 3.

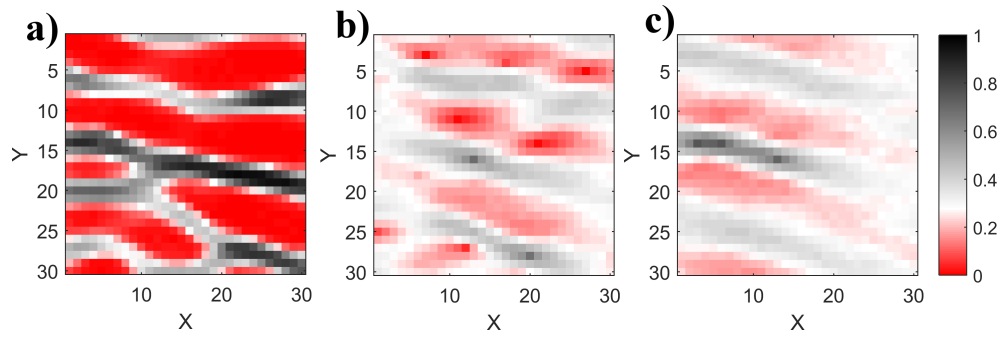


Figure 7: Posterior probability of locating a channel using the DS algorithm with a preferential path, conditional to a) $d1$, b) $d2$, and c) $d3$.

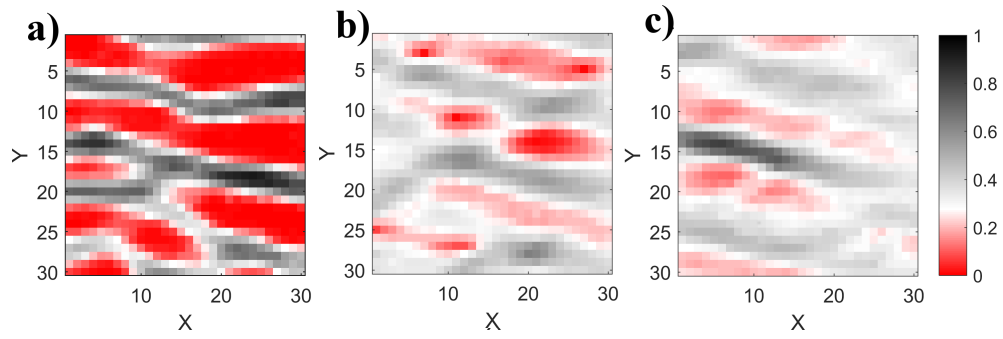


Figure 8: Posterior probability of locating a channel using the SNESIM algorithm with a preferential path, conditional to a) $d1$, b) $d2$, and c) $d3$.

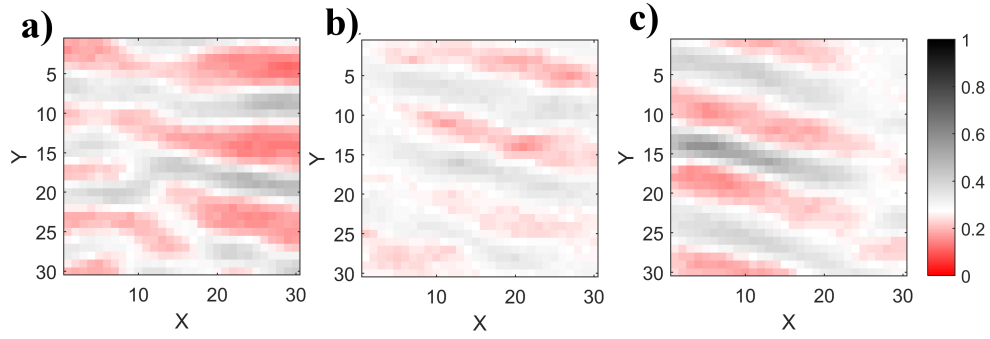


Figure 9: Posterior probability of locating a channel using the DS algorithm using the 3 closest 'soft' data and the 25 closest previously simulated data, with a random path, conditional to a) $d1$, b) $d2$, and c) $d3$.

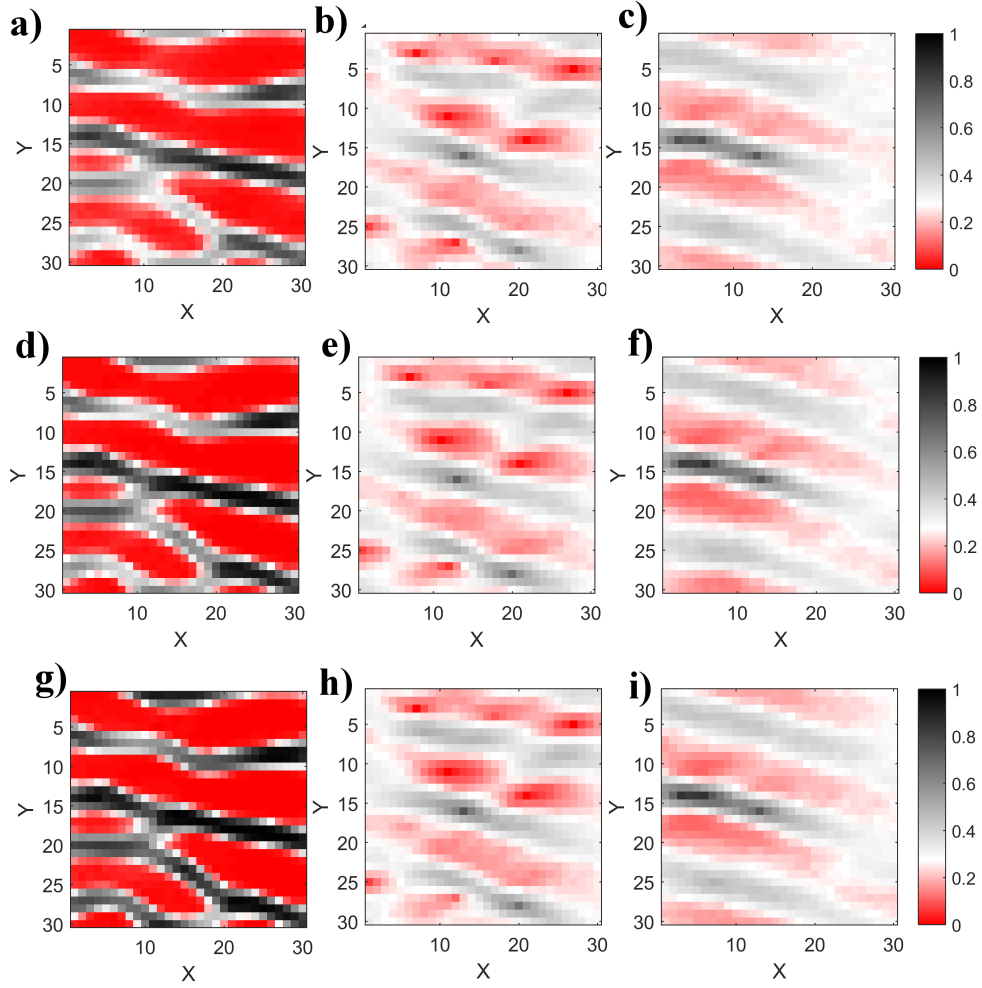


Figure 10: Posterior probability of locating a channel using the DS algorithm using the closest 1 (a,b,c), 3 (d,e,f), and 11 (g,h,i) 'soft'/uncertain data and the 25 closest previously simulated data, with a preferential path, conditional to $d1$ (a,d,f), $d2$ (b,e,g), and $d3$. (c,f,h)

695 **List of Tables**

696	1	The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel)) / L_2(P_{mcmc}(channel))$,	
697		using the GENESIM algorithm and different choices of simula-	
698		tion paths. The left column indicates the conditional data set	
699		considered. Note that the first row for each set of conditional	
700		data, refer to unconditional simulation ($N_{soft} = 0$), for refer-	
701		ence. ‘Markov’ is marked if the Markov property is assumed	
702		such that only co-located data are considered. N_{soft} indicate	
703		the number of closest soft/uncertain data taken into account.	
704		The numbers in parentheses is the simulation time in seconds.	36
705	2	The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel)) / L_2(P_{mcmc}(channel))$,	
706		using the DS algorithm and different choices of simulation	
707		paths. See Table 1 for description.	37
708	3	The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel)) / L_2(P_{mcmc}(channel))$,	
709		using the SNESIM algorithm and different choices of simula-	
710		tion paths. See Table 1 for description.	38

	Markov	N_{soft}	Unilateral	Random	Preferential
$d1$		0	0.78 (42.2 s)	0.77 (59.8 s)	0.77 (59.8 s)
$d1$	*	1	0.63 (40.6 s)	0.70 (59.7 s)	0.43 (39.5 s)
$d1$		1	0.63 (40.7 s)	0.70 (60.0 s)	0.43 (39.7 s)
$d1$		3	0.56 (44.2 s)	0.67 (60.9 s)	0.35 (40.4 s)
$d1$		11	0.47 (46.3 s)	0.65 (64.7 s)	0.25 (46.1 s)
$d2$		0	0.35 (42.3 s)	0.35 (60.4 s)	0.35 (60.6 s)
$d2$	*	1	0.33 (42.7 s)	0.36 (60.7 s)	0.11 (59.9 s)
$d2$		1	0.24 (87.9 s)	0.27 (109.7 s)	0.11 (188.3 s)
$d2$		3	0.19 (167.0 s)	0.25 (145.7 s)	0.11 (189.7 s)
$d2$		11	0.21 (284.4 s)	0.22 (189.6 s)	0.10 (191.1 s)
$d3$		0	0.35 (42.5 s)	0.35 (60.1 s)	0.35 (60.8 s)
$d3$	*	1	0.34 (42.7 s)	0.36 (60.4 s)	0.10 (60.9 s)
$d3$		1	0.24 (186.3 s)	0.26 (152.0 s)	0.10 (179.7 s)
$d3$		3	0.18 (273.7 s)	0.17 (180.7 s)	0.07 (178.6 s)
$d3$		11	0.18 (270.3 s)	0.18 (182.0 s)	0.07 (179.7 s)

Table 1: The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel))/L_2(P_{mcmc}(channel))$, using the GENESIM algorithm and different choices of simulation paths. The left column indicates the conditional data set considered. Note that the first row for each set of conditional data, refer to unconditional simulation ($N_{soft} = 0$), for reference. ‘Markov’ is marked if the Markov property is assumed such that only co-located data are considered. N_{soft} indicate the number of closest soft/uncertain data taken into account. The numbers in parentheses is the simulation time in seconds.

	Markov	N_{soft}	Uni	Random	Preferential
$d1$		0	0.78 (20.8 s)	0.77 (37.3 s)	0.77 (37.4 s)
$d1$	*	1	0.61 (26.1 s)	0.67 (43.9 s)	0.42 (18.3 s)
$d1$		1	0.62 (26.2 s)	0.67 (43.7 s)	0.41 (18.5 s)
$d1$		3	0.52 (51.1 s)	0.65 (74.2 s)	0.34 (19.9 s)
$d1$		11	0.56 (44.2 s)	0.67 (60.9 s)	0.35 (40.4 s)
$d2$		0	0.36 (20.7 s)	0.35 (37.3 s)	0.35 (37.4 s)
$d2$	*	1	0.34 (21.7 s)	0.35 (37.3 s)	0.11 (36.7 s)
$d2$		1	0.20 (30.3 s)	0.28 (53.9 s)	0.10 (45.8 s)
$d2$		3	0.16 (78.5 s)	0.22 (104.2 s)	0.11 (47.4 s)
$d2$		11	0.25 (448.8 s)	0.20 (390.4 s)	0.11 (55.7 s)
$d3$		0	0.36 (19.6 s)	0.34 (38.8 s)	0.34 (37.9 s)
$d3$	*	1	0.33 (19.9 s)	0.36 (39.2 s)	0.09 (38.5 s)
$d3$		1	0.24 (40.0 s)	0.25 (62.8 s)	0.08 (46.9 s)
$d3$		3	0.16 (165.9 s)	0.16 (132.8 s)	0.07 (47.7 s)
$d3$		11	0.17 (163.0 s)	0.16 (135.1 s)	0.07 (47.4 s)

Table 2: The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel))/L_2(P_{mcmc}(channel))$, using the DS algorithm and different choices of simulation paths. See Table 1 for description.

	Markov	N_{soft}	Uni	Random	Preferential
$d1$		0	0.77 (36.0 s)	0.78 (62.6 s)	0.78 (63.2 s)
$d1$	*	1	0.64 (38.2 s)	0.54 (66.1 s)	0.43 (93.1 s)
$d2$		0	0.36 (35.7 s)	0.38 (63.0 s)	0.38 (62.6 s)
$d2$	*	1	0.37 (36.5 s)	0.34 (64.2 s)	0.20 (69.6 s)
$d3$		0	0.34 (36.1 s)	0.36 (63.3 s)	0.36 (63.4 s)
$d3$	*	1	0.36 (36.6 s)	0.35 (63.4 s)	0.16 (58.9 s)

Table 3: The relative L2 norm, $L_2(P_{mcmc}(channel) - P(channel))/L_2(P_{mcmc}(channel))$, using the SNESIM algorithm and different choices of simulation paths. See Table 1 for description.